Japan J. Indust. Appl. Math. (2013) 30:1–19
DOI 10.1007/s13160-012-0089-6

ORIGINAL PAPER

Area 1

Metrics based on average distance between sets

Osamu Fujita

Received: 21 October 2011 / Revised: 26 February 2012 / Published online: 26 June 2012

Abstract This paper presents a distance function between sets based on an average of distances between their elements. The distance function is a metric if the sets are non-empty finite subsets of a metric space. It includes the Jaccard distance as a special case, and can be generalized by using the power mean so as to also include the Hausdorff metric on finite sets. It can be extended to deal with non-null measurable sets, and applied for measuring distances between fuzzy sets and between probability distributions. These distance functions are useful for measuring similarity between data in computer science and information science. In instructional systems design and information retrieval, for example, they are likely to be useful for analyzing and processing text documents that are modeled as hierarchical collections of sets of terms. A distance measure of learners' knowledge is also discussed in connection with quantities of information.

Keywords Metric · Distance between sets · Average distance · Power mean · Hausdorff metric · Information retrieval

Mathematics Subject Classification 51F99 · 68P01 · 68Q01 · 68T01 · 68U01

1 Introduction

A metric defined in general topology [1,2], based on a natural notion of distance between points, is generally extensible to distance between sets or more complex elements. The Hausdorff metric is such a typical one and practically used for image data analysis [3], but it has some problems. In the Euclidean metric on $\mathbb{R}$, for

O. Fujita (✉)
Osaka Kyoiku University, Kashiwara, Osaka 582-8582, Japan
e-mail: fuji@cc.osaka-kyoiku.ac.jp

example, the Hausdorff distance between bounded subsets of $\mathbb{R}$ often depends only on their suprema or infima, no matter how the other elements of the sets are distributed within a certain range, which means it places importance on extremes and disregards the middle parts of the sets. This is a drawback because it is sensitive to noises, errors and outliers in analyzing real world data. There is a need to develop another metric that reflects the overall characteristics of elements of the sets.

In computer science, especially in the fields of artificial intelligence, pattern recognition [4], classification, and information retrieval [5], it is important for data analysis to measure similarity or difference between data such as documents, images and signals. If the data can be represented by vectors, a conventional distance between vectors is a proper measure in their vector space. In practice, however, there are various data that should be dealt with in the form of collections of sets, probability distributions, graph structured data, or collections consisting of more complex data elements. To analyze these data, numerous distance-like functions have been developed [6], like the Mahalanobis distance and the Kullback–Leibler divergence, even though they do not necessarily satisfy symmetry and/or the triangle inequality.

As a true metric, besides the Hausdorff metric, there is another type of distance functions of sets, such as the Jaccard distance, based on the cardinality of the symmetric difference between sets or its variations. However, it measures only the size of the set difference, and takes no account of qualitative differences between individual elements. Thus, both metrics are insufficient to analyze informative data sets in which each element has its own specific meaning.

This paper presents a new distance function between sets based on an average distance. It takes all elements into account. It is a metric if the sets are non-empty finite subsets of a metric space, and includes the Jaccard distance as a special case. By using the power means [7], we obtain generalized forms that also include the Hausdorff metric on finite sets. Extensions of the metric to hierarchical collections of infinite subsets will be useful for treating fuzzy sets and probability distributions, where the distance can be measured not in the function space of the membership or probability density functions but in their domain.

In its application to instructional systems design, for example, the distance function is used for sequencing learning objects such as text documents in order to design textbooks, which are collections of knowledge and concepts, and can be modeled as hierarchical collections of sets of terms. In modeling of knowledge acquisition, a growing space of leaner's knowledge of classification is shown to be evaluated by the distance of partitions of a relevant space in connection with quantities of information. In addition, the feasibility of application to information retrieval and pattern recognition is also discussed.

2 Preliminaries

The metric is extended to various types of generalized metrics. To avoid confusion in terminology, the following definition is used.

Definition 1 (*Metric*) Suppose X is a set and d is a function on $X \times X$ into $\mathbb{R}$. Then d is called a *metric* on X if it satisfies the following conditions, for all $a, b, c \in X$,

M1 $d(a, b) \geq 0$ (non-negativity),
M2 $d(a, a) = 0$,
M3 $d(a, b) = 0 \Rightarrow a = b$,
M4 $d(a, b) = d(b, a)$ (symmetry),
M5 $d(a, b) + d(b, c) \geq d(a, c)$ (triangle inequality).

The set X is called a *metric space* and denoted by (X, d). The function d is called *distance function* or simply *distance*.

The metric is generalized by relaxing the conditions as follows:

- If d satisfies M1, M2, M4 and M5, then it is called a *pseudo-metric*.
- If d satisfies M1, M2, M3 and M5, then it is called a *quasi-metric*.
- If d satisfies M1, M2, M3 and M4, then it is called a *semi-metric*.

This terminology follows [1,2], though the term *"semi-metric"* is sometimes referred to as a synonym of *pseudo-metric* [6].

A set-to-set distance is usually defined as follows (see, e.g., [8]): let A and B be two non-empty subsets of X. For each $x \in X$, the distance from x to A, denoted by $\mathrm{dist}(x, A)$, is defined by the equation

$$\mathrm{dist}(x, A) = \inf\{d(x, a) \mid a \in A\}. \tag{1}$$

This is fundamental not only to the definitions of a boundary point and an open set in metric spaces but also to the generalization of a metric space to *approach space* [9]. Similarly, the distance from A to B can be straightforwardly defined by

$$\mathrm{dist}(A, B) = \inf\{d(a, b) \mid a \in A, b \in B\}. \tag{2}$$

The function dist() is neither a pseudo-metric nor a semi-metric. However, let $\mathscr{S}(X)$ be the collection of all non-empty closed bounded subsets of X. Then, for $A, B \in \mathscr{S}(X)$, the function $h(A, B)$ defined by

$$h(A, B) = \max\{\sup\{\mathrm{dist}(b, A) \mid b \in B\}, \sup\{\mathrm{dist}(a, B) \mid a \in A\}\} \tag{3}$$

is a metric on $\mathscr{S}(X)$, and h is called the *Hausdorff metric*. The collection $\mathscr{S}(X)$ topologized by the metric h is called a *hyperspace* in general topology.

In computer science, data sets are generally discrete and finite. A popular metric is the *Jaccard distance* (or *Tanimoto distance, Marczewski–Steinhaus distance* [6]) that is defined by

$$j(A, B) = \frac{|A \triangle B|}{|A \cup B|}, \tag{4}$$

where $|A|$ is the cardinality of A, and $\triangle$ denotes the symmetric difference: $A \triangle B = (A \setminus B) \cup (B \setminus A)$. In addition, $|A \triangle B|$ is also used as a metric.

In cluster analysis [10], the distance (2) is used as the *minimum distance* between data clusters for single-linkage clustering, and likewise the *maximum distance* is

defined by replacing infimum with maximum for complete-linkage clustering. Moreover, the *group-average distance* (or *average distance, mean distance*) defined as $g(A, B)$ in the following is also typically used for hierarchical clustering. Although these three distance functions are not metrics, the group-average distance plays an important role in this paper.

Lemma 1 *Suppose (X, d) is a non-empty metric space. Let $\mathscr{S}(X)$ denote the collection of all non-empty finite subsets of X. For each A and B in $\mathscr{S}(X)$, define $g(A, B)$ on $\mathscr{S}(X) \times \mathscr{S}(X)$ to be the function*

$$g(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b). \tag{5}$$

Then g satisfies the triangle inequality.

Proof The triangle inequality for d yields $d(a, b) + d(b, c) - d(a, c) \geq 0$ for all $a, b, c \in X$. Then, for all $A, B, C \in \mathscr{S}(X)$, we have

$$\begin{aligned} &g(A, B) + g(B, C) - g(A, C) \\ &= \frac{1}{|A||B||C|} \sum_{a \in A} \sum_{b \in B} \sum_{c \in C} (d(a, b) + d(b, c) - d(a, c)) \geq 0. \end{aligned} \tag{6}$$

□

For ease of notation, let $s(A, B)$ be the sum of all pairwise distances between A and B such that

$$s(A, B) = \sum_{a \in A} \sum_{b \in B} d(a, b), \tag{7}$$

so that $g(A, B) = (|A||B|)^{-1} s(A, B)$. Since d is a metric, we have $s(A, B) \geq 0$, $s(A, B) = s(B, A)$, and $s(\{x\}, \{x\}) = 0$ for all $x \in X$. If $A = \emptyset$ or $B = \emptyset$, then $s(A, B) = 0$ due to the empty sum. If A and B are countable unions of disjoint sets, it can be decomposed as follows:

$$s\left(\bigcup_{i}^{n} A_i, \bigcup_{j}^{m} B_j \right) = \sum_{i}^{n} \sum_{j}^{m} s(A_i, B_j), \tag{8}$$

where $A_i \cap A_j = \emptyset = B_i \cap B_j$ for $i \neq j$. Furthermore, we define $t(A, B, C)$ by the following equation

$$t(A, B, C) = |C| s(A, B) + |A| s(B, C) - |B| s(A, C). \tag{9}$$

It follows from Lemma 1 that $t(A, B, C) \geq 0$ for $A, B, C \in \mathscr{S}(X)$, which is a shorthand notation of the triangle inequality (6).

3 Metric based on average distance

Theorem 1 *Suppose (X, d) is a non-empty metric space. Let $\mathscr{S}(X)$ denote the collection of all non-empty finite subsets of X. For each A and B in $\mathscr{S}(X)$, define $f(A, B)$ on $\mathscr{S}(X) \times \mathscr{S}(X)$ to be the function*

$$f(A, B) = \frac{1}{|A \cup B||A|} \sum_{a \in A} \sum_{b \in B \setminus A} d(a, b) + \frac{1}{|A \cup B||B|} \sum_{a \in A \setminus B} \sum_{b \in B} d(a, b). \quad (10)$$

Then f is a metric on $\mathscr{S}(X)$.

Proof The function f can be rewritten, using s in (7), as

$$f(A, B) = \frac{s(A, B \setminus A)}{|A \cup B||A|} + \frac{s(A \setminus B, B)}{|A \cup B||B|}.$$

It is non-negative and symmetric. If $A = B$, then $s(A, B \setminus A) = s(A, \emptyset) = 0$ and $s(A \setminus B, B) = s(\emptyset, B) = 0$, so that $f(A, B) = 0$. Conversely, if $f(A, B) = 0$, then $s(A, B \setminus A) = 0 = s(A \setminus B, B)$ for $A, B \in \mathscr{S}(X)$. This holds if, and only if, $B \setminus A = \emptyset = A \setminus B$, which implies $B \subseteq A$ and $A \subseteq B$. Then we have $f(A, B) = 0 \Leftrightarrow A = B$.

The triangle inequality is straightforwardly proved to be $f(A, B) + f(B, C) - f(A, C) \geq 0$ by showing that the left-hand terms are transformed into the sum of non-negative terms of s and t in (9). Let $A \cup B \cup C$ be partitioned into five disjoint parts: $\alpha = A \setminus (B \cup C)$, $\beta = B \setminus (A \cup C)$, $\gamma = C \setminus (A \cup B)$, $\zeta = A \cap C \setminus B$, and $\theta = B \setminus \beta = B \cap (A \cup C)$. Then we have

$$\begin{aligned}
&|A||B||C||A \cup B||B \cup C||A \cup C|(f(A, B) + f(B, C) - f(A, C)) \\
&\quad = |B||C|(|\theta \cup C|t(A, B \setminus A, \gamma) + |\alpha|t(A, \beta, \gamma) + |B \cup \zeta|t(A, \beta, C \setminus A)) \\
&\qquad + |A||B|(|A \cup \theta|t(\alpha, B \setminus C, C) + |\gamma|t(\alpha, \beta, C) + |B \cup \zeta|t(A \setminus C, \beta, C)) \\
&\qquad + |A||C|(|A \setminus B|t(\alpha, B, C \setminus B) + |B|t(\alpha, \theta, \gamma) + |C \setminus B|t(A \setminus B, B, \gamma)) \\
&\qquad + |A||C||\theta|(t(\alpha, B, \gamma) + t(\alpha, B, \zeta) + t(\zeta, B, \gamma)) \\
&\qquad + 2|A||C|(|\theta \cup C||A \cup \theta| + |\beta||A \cup C|)s(B, \zeta) \\
&\qquad + 2|A||B||C||B \cup \zeta|s(\beta, A \cap C) \geq 0.
\end{aligned}$$

The details are given in Appendix A. □

The function f in (10) can be rewritten, using g in (5), as

$$f(A, B) = \frac{|A \setminus B|}{|A \cup B|} g(A \setminus B, B) + \frac{|B \setminus A|}{|A \cup B|} g(B \setminus A, A). \quad (11)$$

In $(\mathscr{S}(X), f)$, for all $a, b \in X$, we have $f(\{a\}, \{b\}) = d(a, b)$ so that $\{\{x\} \mid x \in X\}$ is an isometric copy of X. If $A \cap B = \emptyset$, then $f(A, B) = g(A, B)$. If d is a pseudo-metric, then so is f.

Example 1 If d is the discrete metric, where $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ otherwise, then $f(A, B)$ is equal to the Jaccard distance (4).

Corollary 1 *Suppose (X, d) is a non-empty metric space. Let $\mathscr{S}(X)$ denote the collection of all non-empty finite subsets of X. For each A and B in $\mathscr{S}(X)$, define $e(A, B)$ on $\mathscr{S}(X) \times \mathscr{S}(X)$ to be the function*

$$e(A, B) = \frac{1}{|A||B|}\left(\sum_{a\in A}\sum_{b\in B} d(a, b) - \sum_{a\in A\cap B}\sum_{b\in A\cap B} d(a, b)\right).$$

Then e is a semi-metric on $\mathscr{S}(X)$.

Proof Let $e(A, B)$ be rewritten as

$$e(A, B) = \frac{1}{|A||B|}\left(s(A \setminus B, B \setminus A) + s(A \cap B, B \setminus A) + s(A \setminus B, A \cap B)\right).$$

In a similar manner to the proof of Theorem 1, it can be proved that the conditions from M1 to M4, except for M5 (triangle inequality), are satisfied. □

Remark 1 It is noted that the triangle inequality $e(A, B) + e(B, C) \geq e(A, C)$ holds if $|\delta||\varepsilon||\eta| = 0$, where $\delta = A \cap B \setminus C$, $\varepsilon = B \cap C \setminus A$ and $\eta = A \cap B \cap C$. Otherwise, for example, if $A = \delta \cup \eta$, $B = \delta \cup \eta \cup \varepsilon$ and $C = \eta \cup \varepsilon$ for non-empty δ, ε and η, then we have

$$\begin{aligned}&|A||B||C|(e(A, B) + e(B, C) - e(A, C))\\ &\quad = |\delta|s(\delta, \varepsilon) - |\varepsilon|s(\delta, \eta) - |\delta|s(\eta, \varepsilon) = -t(\delta, \eta, \varepsilon) \leq 0,\end{aligned}$$

so that the condition M5 is not generally satisfied.

4 Extensions

This section discusses future directions for generalization of the average distance based on the power mean and extensions to metrics on collections of infinite sets.

4.1 Generalization based on the power mean

The distance function (10) can be unified with the Hausdorff metric for finite sets by using the power mean. To simplify expressions, we use the following notation. Let $M_p^{(i)}(x \in A, \psi, w)$ be an extended weighted-power-mean of $\psi(x)$ such that

$$M_p^{(1)}(x \in A, \psi, w) = \left(\frac{1}{\sum_{x\in A} w(x)}\sum_{x\in A} w(x)(\psi(x))^p\right)^{1/p}, \tag{12}$$

and its variation using the exponential transform of ψ,

$$M_p^{(0)}(x \in A, \psi, w) = \frac{1}{p} \ln\left(\frac{1}{\sum_{x\in A} w(x)} \sum_{x\in A} w(x) \exp(p\psi(x)) \right), \tag{13}$$

where $i \in \{0, 1\}$ indicates one of the two types (12) and (13), p is an extended real number, ψ is a non-negative function of $x \in A$, and w is a weight such that $w(x) \in (0, 1]$ for each x and $\sum_{x\in A} w(x) > 0$. In addition, let $M_p^{(i)}(x \in A, \psi)$ denote the abbreviation of the equal weight case $M_p^{(i)}(x \in A, \psi, 1_A(x))$, where $1_A(x)$ is the indicator function defined by $1_A(x) = 1$ for $x \in A$ and $1_A(x) = 0$ for $x \notin A$. If there exists $x \in A$ such that $\psi(x) = 0$ for $p < 0$ in (12), then we define $M_p^{(1)} = 0$, which is consistent with taking the limit $\psi(x) \to 0^+$, though such a case is undefined in the conventional power mean to avoid division by zero.

The power mean includes various types of means [7], which are parameterized by p. By taking limits also for $p = 0, \pm\infty$, we have the following:

$$\begin{aligned}
M_i^{(i)}(x \in A, \psi(x), 1_A(x)) &= M_i^{(i)}(x \in A, \psi(x)) = \frac{1}{|A|} \sum_{x\in A} \psi(x), \\
M_\infty^{(i)}(x \in A, \psi(x), w(x)) &= M_\infty^{(i)}(x \in A, \psi(x)) = \max\{\psi(x) \mid x \in A\}, \\
M_{-\infty}^{(i)}(x \in A, \psi(x), w(x)) &= M_{-\infty}^{(i)}(x \in A, \psi(x)) = \min\{\psi(x) \mid x \in A\}.
\end{aligned}$$

Both $M_1^{(1)}$ and $M_0^{(0)}$ give the same arithmetic mean, whereas $M_0^{(1)}$ gives the geometric mean, and neither $M_\infty^{(i)}$ nor $M_{-\infty}^{(i)}$ depends on $w(x)$.

There are some forms of function composition of $M_p^{(i)}$ that include the distance function (10) as a special case, for example, as follows:

$$\begin{aligned}
u_{p,q}^{(i,j)}(A, B) = M_p^{(i)} \Big(x \in A \cup B, \Big(& 1_{B\setminus A}(x)\, M_q^{(j)}(y \in A, d(x, y)) \\
& + 1_{A\setminus B}(x)\, M_q^{(j)}(y \in B, d(x, y)) \Big) \Big),
\end{aligned} \tag{14}$$

$$v_{r,p,q}^{(k,i,j)}(A, B) = M_r^{(k)} \left(S \in \{A, B\}, M_p^{(i)} \left(x \in A \cup B, 1_{S^{\complement}}(x)\, M_q^{(j)}(y \in S, d(x, y)) \right) \right), \tag{15}$$

where $i, j, k \in \{0, 1\}$ and $S^{\complement}$ is the complement of S.

In addition, let w be extended to $w \in [0, 1]$ to include zero, though a weight for at least one summand must still be positive. Furthermore, it is assumed that $0 \cdot \infty = 0$, in order to ensure $0 \cdot 0^p = 0$ for $p < 0$ in $M_p^{(1)}$, so that the zero-weight can be used for excluding terms from averaging even if $\psi(x) = 0$ (i.e., distance zero) in the terms. Then, the function (14) can be simply expressed as

$$u_{p,q}^{(i,j)}(A, B) = M_p^{(i)} \left(x \in A \cup B, M_q^{(j)}(y \in A \cup B, d(x, y), w(x, y)) \right), \tag{16}$$

by using the weight function defined by

$$\begin{aligned} w(x, y) &= [x \in A \setminus B][y \in B] + [x \in A \cap B][x = y] + [x \in B \setminus A][y \in A] \\ &= 1_{A \setminus B}(x)\, 1_B(y) + 1_{A \cap B}(x)[x = y] + 1_{B \setminus A}(x)\, 1_A(y), \end{aligned}$$

where $[\cdot]$ denotes the Iverson bracket, that is a quantity defined to be 1 whenever the statement within the brackets is true, and 0 otherwise.

The distance function f in (10) and the Hausdorff metric (3) are expressed by

$$\begin{aligned} f(A, B) &= u_{i,j}^{(i,j)}(A, B) = 2\, v_{k,i,j}^{(k,i,j)}(A, B), \\ h(A, B) &= u_{\infty,-\infty}^{(i,j)}(A, B) = v_{\infty,\infty,-\infty}^{(k,i,j)}(A, B), \end{aligned}$$

respectively, where $i, j, k \in \{0, 1\}$.

Although it is unclear, at present, what conditions on the parameters i, j, k, p, q, and r are necessary for (14) and (15) to be metrics, these generalized forms are capable of generating various distance functions in fact as follows.

Example 2 The log-exp types $u_{p,q}^{(0,0)}(A, B)$ and $v_{r,p,q}^{(0,0,0)}(A, B)$ are written as

$$\begin{aligned} u_{p,q}^{(0,0)}(A, B) = \frac{1}{p} \ln \Bigg(& \frac{1}{|A \cup B|} \sum_{x \in A \cup B} \Bigg(\frac{1_{B \setminus A}(x)}{|A|} \sum_{y \in A} e^{qd(x,y)} \\ & + 1_{A \cap B}(x) + \frac{1_{A \setminus B}(x)}{|B|} \sum_{y \in B} e^{qd(x,y)} \Bigg)^{p/q} \Bigg), \end{aligned}$$

$$\begin{aligned} v_{r,p,q}^{(0,0,0)}(A, B) = \frac{1}{r} \ln \Bigg(& \frac{1}{2} \Bigg(\frac{1}{|A \cup B|} \sum_{x \in B \setminus A} \Bigg(\frac{1}{|A|} \sum_{y \in A} e^{qd(x,y)} \Bigg)^{p/q} + \frac{|A|}{|A \cup B|} \Bigg)^{r/p} \\ & + \frac{1}{2} \Bigg(\frac{1}{|A \cup B|} \sum_{x \in A \setminus B} \Bigg(\frac{1}{|B|} \sum_{y \in B} e^{qd(x,y)} \Bigg)^{p/q} + \frac{|B|}{|A \cup B|} \Bigg)^{r/p} \Bigg). \end{aligned}$$

If d is the discrete metric multiplied by a positive constant λ, then we have

$$u_{p,q}^{(0,0)}(A, B) = \frac{1}{p} \ln \left(\frac{e^{p\lambda} |A \triangle B| + |A \cap B|}{|A \cup B|} \right), \tag{17}$$

$$v_{0,p,q}^{(0,0,0)}(A, B) = \frac{1}{2p} \ln \left(\left(\frac{e^{p\lambda} |B \setminus A| + |A|}{|A \cup B|} \right) \left(\frac{e^{p\lambda} |A \setminus B| + |B|}{|A \cup B|} \right) \right). \tag{18}$$

The functions (17) and (18) are metrics for $p > 0$ and $p < 0$, respectively. The proofs of each triangle inequality are outlined in Appendix B. If $p = 0$, then it is the same

situation as Example 1. By taking the limit for $p = 0$, we can see that both functions are equal to the Jaccard distance (4) except for the coefficient λ.

4.2 Hierarchical metric spaces

Suppose that (X, d) is a metric space and $\mathscr{S}(X)$ is the collection of all non-empty finite subsets of X. Let k be a non-negative integer, let $\mathscr{S}_{k+1}(X)$ denote the collection of all non-empty finite subsets of $\mathscr{S}_k(X)$, and let f_k be a metric on $\mathscr{S}_k(X)$, where $\mathscr{S}_1(X)$, $\mathscr{S}_0(X)$, f_1 and f_0 correspond to, respectively, $\mathscr{S}(X)$, X, f, and d in Theorem 1. For $k > 1$, in much the same way, for each $\mathscr{A}$ and $\mathscr{B}$ in $\mathscr{S}_k(X)$, the function $f_k(\mathscr{A}, \mathscr{B})$ can be defined by

$$f_k(\mathscr{A}, \mathscr{B}) = \frac{\sum_{A\in\mathscr{A}} \sum_{B\in\mathscr{B}\setminus\mathscr{A}} f_{k-1}(A, B)}{|\mathscr{A} \cup \mathscr{B}||\mathscr{A}|} + \frac{\sum_{A\in\mathscr{A}\setminus\mathscr{B}} \sum_{B\in\mathscr{B}} f_{k-1}(A, B)}{|\mathscr{A} \cup \mathscr{B}||\mathscr{B}|}, \tag{19}$$

which generates a metric space $(\mathscr{S}_k(X), f_k)$ based on $(\mathscr{S}_{k-1}(X), f_{k-1})$. This metric will be useful for constructing hierarchical hyperspaces.

4.3 Duality

There is a kind of duality between sets and elements with respect to their distance functions. For example, we can define the functions D and d symmetrically as follows:

$$D(A, B) = |\{a \mid a \in A\} \triangle \{b \mid b \in B\}| = |A \triangle B|, \tag{20}$$
$$d(a, b) = |\{A \mid a \in A\} \triangle \{B \mid b \in B\}| = |\mathscr{C}(a) \triangle \mathscr{C}(b)|, \tag{21}$$

where $\mathscr{C}(a) = \{A \mid a \in A\}$. The set-to-set distance D in (20) is a metric due to the axiom of extensionality, and the element-to-element distance d in (21) is a pseudo-metric.

According to Theorem 1, D can be defined by f in (10), instead of (20), so that we have

$$D(A, B) = f(A, B),$$
$$d(a, b) = |\mathscr{C}(a) \triangle \mathscr{C}(b)|.$$

In this case, D is a pseudo-metric, depending on d, and there exist a condition that satisfy $d(a, b) = f(\bigcap \mathscr{C}(a), \bigcap \mathscr{C}(b))$. Furthermore, in the situation of Sect. 4.2, let (X, d) be a metric space, let $\mathscr{S}_1(X)$ be the collection of all non-empty finite subsets of X, and let $\mathscr{C}(a) = \{A \in \mathscr{S}_1(X) \mid a \in A\}$ be an element of $\mathscr{S}_2(X)$. If d is the

discrete metric, then we have

$$D(A, B) = f_1(A, B),$$
$$f_2(\mathscr{C}(a), \mathscr{C}(b)) = \kappa d(a, b),$$

where κ is a certain positive real number such that $\kappa < 1$. If there exists d such that $d(a, b) = f_2(\mathscr{C}(a), \mathscr{C}(b))$, then the isometric copy of (X, d) is contained in $(\mathscr{S}_2(X), f_2)$ and d can be regarded as the function of D. In general, there possibly exist D and d that are formally expressed by

$$D(A, B) = F(\{d(a, b) \mid a \in A, b \in B\}), \tag{22}$$
$$d(a, b) = G(\{D(A, B) \mid a \in A, b \in B\}), \tag{23}$$

where F and G may be such a generalized function given in (16). It is interesting to consider whether F and G really exist and what features they have. In numerical analysis, D and d that are consistent with each other will be obtained by the iterative computation of (22) for all $A, B \in \mathscr{S}_1(X)$ and (23) for all $a, b \in X$, starting with an initial metric space (X, d), if each converges to a non-trivial function.

4.4 Generalized metrics

The group average distance $g(A, B)$ in (5) can be regarded as a generalized metric that satisfies conditions M1 (non-negativity), M4 (symmetry) and M5 (triangle inequality) in Definition 1. In conventional topology, there has been no such generalization by dropping both conditions M2 and M3, which are usually combined together into the single axiom $d(a, b) = 0 \Leftrightarrow a = b$ (identity, reflexivity, or coincidence). Although M3 can be dropped for the pseudo-metric so as to allow $d(a, b) = 0$ for $a \neq b$, the self-distance $d(a, a) = 0$ (M2) seems to be indispensable in point-set topology where the element is a point having no size. An exception is a *partial metric* [11] that is defined to satisfy M1, M3, M4 and, instead of M5, the following partial metric triangularity,

$$d(a, b) + d(b, c) \geq d(a, c) + d(b, b). \tag{24}$$

In computer science or information science, the element of data sets is not merely a simple point. It may have rich contents inside. Some elements may have internal structures which cause non-zero self-distance, and some elements may have different properties each other, even though they are indiscernible from a metric point of view. The concept of distance can be used for measuring not only the difference between objects but also the cost of moving or the energy of transition between states. This is the reason why the generalization toward non-zero self-distance is worth considering. If the triangle inequality holds, it provides an upper and lower bound for them. The group average distance $g(A, B)$ can be such a typical one, and that it is simpler and more natural than the metric $f(A, B)$ in (10).

Incidentally, the function $g(A, B)$ is not a partial metric because it does not satisfy (24). On the other hand, $f(A, B)$ gives an approach to an instance of partial metrics from a special case of (18) in Example 2. By taking the limit as $\lambda \to \infty$ for $p < 0$ and multiplying a positive constant, for non-empty finite sets A and B, we have the following metric,

$$D_\nu(A, B) = \log |A \cup B| - \nu \log(|A||B|),$$

where $\nu = 1/2$. Its triangle inequality is equivalent to $|A \cup B||B \cup C| \geq |A \cup C||B|$. This suggests, for $\nu \in [0, 1/2)$, D_ν is a partial metric on a collection of non-empty finite sets.

4.5 Extension to infinite sets

If $\mathscr{S}(X)$ is the collection of all non-null measurable subsets of (X, d), and d is Lebesgue integrable on each element of $\mathscr{S}(X)$, then the group average distance $g(A, B)$, for $A, B \in \mathscr{S}(X)$, can be defined by

$$g(A, B) = \frac{1}{\mu(A)\mu(B)} \int_A \left(\int_B d(x, y) \mathrm{d}\mu(y) \right) \mathrm{d}\mu(x) \tag{25}$$

where $x \in A$, $y \in B$, and μ is a measure on X, and then the distance function (11) can be extended to

$$f(A, B) = \frac{\mu(A \setminus B)}{\mu(A \cup B)} g(A \setminus B, B) + \frac{\mu(B \setminus A)}{\mu(A \cup B)} g(B \setminus A, A). \tag{26}$$

If d is the discrete metric, then (26) is equal to the *Steinhaus distance* [6].

Example 3 Let $(\mathbb{R}, d)$ be a metric space and let $d(x, y) = |x - y|$. For two intervals A and B, the distance function (26) can be expressed as

$$f(A, B) = \frac{|\sup(A) - \sup(B)| + |\inf(A) - \inf(B)|}{2} - \frac{|\sup(A) - \sup(B)||\inf(A) - \inf(B)|}{\sup(A \cup B) - \inf(A \cup B)} [(A \subset B) \vee (A \supset B)]$$

If $A \not\subseteq B$ and $A \not\supseteq B$ (i.e., $[(A \subset B) \vee (A \supset B)] = 0$), then $f(A, B)$ is equal to the distance between the centers of A and B. This is consistent with an intuitive notion of the distance between balls in this $(\mathbb{R}, d)$.

If $\mathscr{S}(X)$ is the collection of all non-empty, countably infinite subsets (measure-zero sets) of X, then $g(A, B)$ and $f(A, B)$ should be defined by taking limits in (5) and (11), provided that both have definite values. In order to determine the average distance, we have to define a proper condition, which should be said to be "averageable".

The average distance will strongly depend on accumulation points in A and B, and it will require additional assumptions on the difference of the strength between the accumulation points. This requirement is closely related to a "relative measure" that is needed to obtain the ratio of the cardinality of an infinite set to the cardinality of its superset in (11). In conventional measure theory, however, any set of cardinality $\aleph_0$ is a null set having measure zero so that both counting measure and Lebesgue measure are useless for computing the ratio. It is necessary to use another measure. A feasible solution is discussed in the following section.

4.6 Estimation by sampling

In application to computational data analysis, statistical estimation by sampling is very useful for obtaining the approximate value of $f(A, B)$ when the size of the sets is very large. According to the law of large numbers, if enough sample elements are selected randomly, an average generated by those samples should approximate the average of the total population. The procedure is as follows:

1. Choose a superset P of $A \cup B$ as a population such that $P \supseteq A \cup B$.
2. Select a finite subset S of P as a sample obtained by random sampling.
3. Let $S_A = S \cap A$ and $S_B = S \cap B$. Then, compute $f(S_A, S_B)$ for approximation of $f(A, B)$.

The sampling process and its randomness are crucial for efficiently estimating a good approximation. Some useful hints could be found in various sampling techniques developed for Monte Carlo methods [12]. In most cases, sampling error is expected to decrease as the sample size increases, except for situations where the distribution of d has no mean (e.g., Cauchy distribution).

The notion of sampling suggests an intuitive measure to define a relative measure on a σ-algebra over a set X, which could be called "sample counting measure". Suppose A and B are subsets of X. Let $\rho(A : B)$ be the ratio of the cardinality of A to the cardinality of B, let P be a superset of $A \cup B$, and let S_n be a non-empty finite subset of P such that $S_n = \bigcup_{i=1}^{n} Y_i$ where Y_i is the ith non-empty sample randomly selected from P. Then, the ratio $\rho(A : B)$ can be determined by taking a limit of n as it approaches to ∞ as follows:

$$\rho(A : B) = \lim_{n \to \infty} \frac{|S_n \cap A|}{|S_n \cap B|},$$

if there exist such a limit and a random choice function that performs random sampling. Otherwise, instead of random sampling, systematic sampling could be available if the elements of X are supposed to be distributed with uniform density in its measurable metric space. For example, suppose there exists a finite partition of P where every part has an almost equal diameter. It seems better for S_n to have exactly one element with each of the parts.

4.7 Metrics for fuzzy sets and probability distributions

A fuzzy set can be represented by a collection of crisp sets so that the distance between fuzzy sets can be defined by the distance between the collections of such crisp sets. Let A be a fuzzy set: $A = \{(x, m_A(x)) \mid x \in X\}$, where $m_A(x)$ is a membership function, and let A_α be a crisp set called an $\alpha - lebel$ set [13] such that $A_\alpha = \{x \in X \mid m_A(x) \geq \alpha\}$. Then, A can be represented by the following set of ordered pairs: $\mathscr{C}(A) = \{(A_\alpha, \alpha) \mid \alpha \in (0, 1]\}$. The distance between two fuzzy sets A and B can be defined as $f(\mathscr{C}(A), \mathscr{C}(B))$, where the distance between (A_α, α) and (B_β, β) can be defined as $f(A_\alpha, B_\beta) + d(\alpha, \beta)$, for example. This notion is also applicable to the distance between probability distributions, where probability density functions are used instead of the membership function.

5 Application

This section discusses the application of the distance function f defined in (10) and its extensions in the field of computer science, where the distance is widely used for measuring the similarity/dissimilarity of data. Some of the advantages of using f for measuring distance between sets are: f is a metric for all nonempty finite subsets of a metric space so that all measurements are consistent with each other; the average is the fundamental statistic of data, which can be estimated with sampling; and it is less sensitive to noises, errors and outliers of data than such a max-of-min distance as the Hausdorff metric.

The following topics deal mainly with text documents, which can be applied to various kinds of data such as audio, images, video, and any other information. In practice, there are a variety of distance measures developed and used. In addition to them, f provides some advanced options. However, the practical advantages and disadvantages of using them generally vary with each case, where even quasi-metrics or semi-metrics can be a reasonable option, so that its performance evaluation is beyond the scope of this paper.

5.1 Instructional systems design

The distance between text documents can be used to sequence learning objects, which is useful especially for e-learning systems that enable to automatically and dynamically compose personalized lessons for an individual learner [14]. For example, a textbook is a collection of the knowledge and concepts of a specific subject. The knowledge that is describable and computationally manageable can be represented as hierarchical or well-founded sets of text data, which is structured as chapters, sections and paragraphs. In each set of each level, its elements should be well organized to have smooth transitions with adjacent elements for readability. The smoothness of the transition can be evaluated as the distance between the text data.

Text document data can be represented as sets of terms, which may be not only words or phrases but also images, graphs and equations, and that they might have more complex data structures in practice. Each term has its own specific meaning and is

semantically equivalent to proper text data as described in dictionaries. There can be a distance between terms, as well as between text documents, for instance, which may be defined by (21) for a given collection of articles such as encyclopedia, dictionary, and thesaurus.

Let d be the distance between terms and let S_i be a set of terms representing the ith section. For a given d and $\{S_i\}$, the distance between S_i and S_j can be defined by f in (10). Then the subsequent section should be arranged so as to minimize $f(S_i, S_{i+1})$, though some exceptions may be allowed to abandon local optimum in order to minimize the total sum of N sections: $\sum_{i=1}^{N-1} f(S_i, S_{i+1})$.

Generally, data transition path analysis is applicable to various kinds of information processing that transforms source data into target or desired data via intermediate states. In the case that there are multiple paths between the data and the distance between them is defined to imply the cost of data processing, f gives an average cost of them and so it can be used for estimating and optimizing the actual total cost of data processing.

5.2 Learner model

Knowledge acquisition and learning processes can be modeled and evaluated with the distance between sets. The sequence of S_i mentioned above is represented not only as a trajectory in a space of teacher's knowledge, but also as a growing space of learner's knowledge. As the complexity of the space increases, the distances between sets of the knowledge increase, as well as the distances between their elements, if they are defined as (21).

Let us consider the knowledge that classifies elements of a universal set U into a number of classes. A subset S of U represents a fundamental piece of the knowledge that dichotomizes U into S and $S^{\complement}$. A learner acquires a collection of S from a training collection $\mathscr{T}$ of a teacher. The learner can increase the collection, if he/she has the intelligence to infer logical combinations of the acquired subsets. Suppose that, consequently, the learner becomes to have the knowledge that enables to recognize N classes, which is represented by a partition $\mathscr{P} = \{P_1, P_2, \ldots, P_N\}$ of U. Let $P(a)$ be a part that contain the element a, i.e., $a \in P(a) \in \mathscr{P}$, and let $d(a, b)$ be a pseudo-metric defined as

$$d(a,b) = \begin{cases} \log(|U|^2/|P(a)||P(b)|) & \text{if } P(a) \neq P(b), \\ 0 & \text{if } P(a) = P(b), \end{cases} \tag{27}$$

where the elements in the same part are indiscernible for the learner, due to lack of the knowledge.

The distance of the partition $\mathscr{P}$ from the initial state of $\mathscr{P}_0 = \{U\}$ indicates a degree of the knowledge accumulation. Using (27), for example, we have

$$f_2(\mathscr{P}_0, \mathscr{P}) = \frac{1}{N} \sum_{i=1}^{N} f_1(U, P_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{|U||P_i|}\sum_{a\in U\setminus P_i}\sum_{b\in P_i} d(a,b)\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N}\log\frac{|U|}{|P_i|} + \frac{N-2}{N}\sum_{i=1}^{N}\frac{|P_i|}{|U|}\log\frac{|U|}{|P_i|}. \tag{28}$$

Incidentally, considering that $\log(|A\cup B|^2/|A||B|)$ is a metric for $A, B \subseteq U$, the first term of the right hand side of (28) can, by itself, be regarded as a distance between them. Furthermore, if U is regarded as the sample space of a probability space and its all elements have an equal probability, the first term is the mean of the self-information of the parts, and the second term implies entropy. The first term is minimum, whereas the second is maximum, when all $|P_i|$ are equal.

It is also important to consider the distance between learners and teachers. In practice, however, acquired subsets of learners may alter from their original in $\mathscr{T}$ or contain elements that are unknown to the learners. To optimize learning processes for these cases, different types of learner models will be needed, which are subjects for future research.

5.3 Information retrieval

In document retrieval systems [5], the similarity measure of documents is crucial for finding relevant documents to user queries and ranking them. In the vector space model using term frequency-inverse document frequency (tf-idf) [15], for example, text documents are represented as vectors where each component corresponds to the tf-idf value of a particular term, and the similarity is measured by the cosine of the angle between the vectors. The documents are ranked according to the similarity to a query, though the relevant documents do not always have a high similarity score, whereas non-relevant documents sometimes have a higher score. These inconsistencies in the ranking should be due to the crudity of the vector representation and the similarity measure.

The performance of information retrieval systems is commonly evaluated by measuring precision and recall. Let S be a set of retrieved documents and let T be a set of relevant documents. Then, precision and recall are defined, respectively, as $|S\cap T|/|S|$ and $|S\cap T|/|T|$. Furthermore, the F-measure defined as their harmonic mean $2|S\cap T|/(|S|+|T|)$ is also commonly used for optimizing the performance. To maximize the F-measure, i.e., to exclude non-relevant documents out of the ranking and/or to give a higher rank to the relevant documents, it is necessary to modify the details of the model. However, the F-measure has a drawback that it is insensitive to rearranging the ranking within $S\cap T$ or $S\setminus T$.

For improving the model, $f(S,T)$ is more useful than the F-measure, because it can be made differentiable with respect to model parameters that are included in document vectors, query vectors and their distance function, provided that the distance function is defined consistently with the similarity measure. The typical process of model tuning is as follows:

1. let U be a set of all documents, which is assumed to also contain known and unknown queries;
2. make a metric space (U, d), where d is defined as the distance between documents;
3. prepare a training set $\mathscr{T}$ that each element is the pair of a set of queries $Q \subset U$ and its relevant documents $T \subset U$;
4. modify model parameters related to $(Q, T) \in \mathscr{T}$ so as to minimize $\sum_{(Q,T)\in\mathscr{T}} f(Q, T)$;
5. retrieve a set of documents $S(Q, \hat{r})$ where $\hat{r} = \arg\min f(S(Q, r), T)$ and $S(Q, r) = \{x \in U \mid f(Q, \{x\}) < r\}$;
6. modify model parameters related to $S(Q, \hat{r}) \setminus T$ so as to increase $\sum_{(Q,T)\in\mathscr{T}} f(Q, S(Q, \hat{r}) \setminus T)$;
7. modify model parameters related to $T \setminus S(Q, \hat{r})$ so as to decrease $\sum_{(Q,T)\in\mathscr{T}} f(Q, T \setminus S(Q, \hat{r}))$;
8. repeat the steps from 4 to 7 until $\sum_{(Q,T)\in\mathscr{T}} f(S(Q, \hat{r}), T)$ decreases to less than a desired value.

Generally, the performance measure is not limited to $f(S, T)$ for this purpose. As to the model itself, instead of vector representation, set theoretic and probabilistic models can also be used. The distance d may be defined by f as discussed in Sect. 5.1. In Step 2, (U, d) may be a generalized metric space that d is a generalized metric such as pseudo-, quasi-, and semi-metric. To optimize the model as a whole, it is important to prepare a large training set that contains a wide variety of queries, taking all unknown and unpredictable queries into account. In Step 3, Q is usually a singleton but may have more than two queries for searching a single topic, which occurs by adding queries for refining search results. Step 4 and 7 are overlapped and so Step 7 can be removed, or Step 4 may be limited to modify only Q.

5.4 Pattern recognition

The distance function between documents defined by f, as mentioned above, can also be used for document classification. In the k-nearest neighbor algorithm (k-NN) [16], an unlabeled document is classified by assigning the label which is most frequent among the k nearest neighbors of labeled documents, which are selected in much the same way as top k rank documents obtained in information retrieval.

On the other hand, in conventional methods for pattern recognition [4], mostly input data are transformed into feature vectors and classified in their vector space. For example, the support vector machine (SVM) [17] is a state-of-the-art binary classifier that separates feature vectors with an optimized hyperplane with respect to the distance to its closest support vectors, and is extended to a nonlinear classifier by using the kernel trick [18]. In order to take this approach, though SVM is neither simpler nor much better than k-NN for multiclass classification, it is necessary for the distance f to find an isometry to an appropriate vector space.

Probabilistic models are also widely used for machine learning, and often achieve good performance due to optimal estimation and inference procedures based on statistics and probability theory. A drawback of such models is the loss of information due

to the assumption of random variables that are intrinsically non-random, as used in modeling text documents. To further improve performance, it is important to consider the properties of the domain of probability functions, where the distance measure is likely to provide some useful information.

6 Concluding remarks

We have found that, for a metric space (X, d), there exists a distance function between non-empty finite subsets of X that is a metric based on the average distance of d. The distance function (10) in Theorem 1 is the most typical one, which includes the Jaccard distance as a special case where d is the discrete metric. Its extensions based on the power mean are useful to develop generalized forms that also include the Hausdorff metric on finite sets and the other various distance functions. Furthermore, the extensions to infinite subsets of X will provide metrics for measuring dissimilarity of fuzzy sets and probability distributions. These functions will be useful for measuring similarity/dissimilarity of data such as text documents in the field of computer science and information science, especially for application to instructional systems design and information retrieval.

Appendix A: Triangle inequality in Theorem 1

The triangle inequality for

$$f(A, B) = (|A \cup B||A|)^{-1} s(A, B \setminus A) + (|A \cup B||B|)^{-1} s(A \setminus B, B)$$

can be proved by showing the following inequality:

$$|A||B||C||A \cup B||B \cup C||A \cup C|(f(A, B) + f(B, C) - f(A, C)) \geq 0.$$

Let $A \cup B \cup C$ be partitioned into the following seven disjoint parts:

$$\alpha = A \setminus (B \cup C), \quad \beta = B \setminus (A \cup C), \quad \gamma = C \setminus (A \cup B),$$
$$\delta = A \cap B \setminus C, \quad \varepsilon = B \cap C \setminus A, \quad \zeta = C \cap A \setminus B, \quad \eta = A \cap B \cap C,$$

and let $\theta = B \setminus \beta = B \cap (A \cup C)$, so that we have

$$\begin{aligned} A &= \alpha \cup \delta \cup \zeta \cup \eta, \quad |A| = |\alpha| + |\delta| + |\zeta| + |\eta|, \\ B &= \beta \cup \delta \cup \varepsilon \cup \eta, \quad |B| = |\beta| + |\delta| + |\varepsilon| + |\eta|, \\ C &= \gamma \cup \varepsilon \cup \zeta \cup \eta, \quad |C| = |\gamma| + |\varepsilon| + |\zeta| + |\eta|, \\ \theta &= \delta \cup \varepsilon \cup \eta, \quad |\theta| = |\delta| + |\varepsilon| + |\eta|. \end{aligned}$$

Taking account of (8) and (9), we have

$$
\begin{aligned}
&|A||B||C||A\cup B||B\cup C||A\cup C|(f(A,B)+f(B,C)-f(A,C))\\
&\quad=|B||C||B\cup C||A\cup C|s(A,B\setminus A)+|A||C||B\cup C||A\cup C|s(A\setminus B,B)\\
&\qquad+|A||C||A\cup B||A\cup C|s(B,C\setminus B)+|A||B||A\cup B||A\cup C|s(B\setminus C,C)\\
&\qquad-|B||C||A\cup B||B\cup C|s(A,C\setminus A)-|A||B||A\cup B||B\cup C|s(A\setminus C,C)\\
&\quad=|B||C|(|\gamma||\delta\cup C|s(A,B\setminus A)+(|\gamma||\alpha|+|B\cup\zeta||C\setminus A|)s(A,\beta)\\
&\qquad+|B\cup\zeta||A|(s(A\setminus C,\beta)+s(A\cap C,\beta))+(|\beta||\gamma|+|B\cup C||A\cup\varepsilon|)s(A,\varepsilon))\\
&\qquad+|A||C|((|\gamma||A\cup C|+|\zeta||\gamma|+|\zeta||A\cup\varepsilon|)s(\alpha,B)\\
&\qquad+|B||A\cup\varepsilon|(s(\alpha,B\setminus C)+s(\alpha,B\cap C))+|B||\gamma|(s(\alpha,\beta)+s(\alpha,\theta))\\
&\qquad+(|\delta\cup C||\gamma|+|\delta\cup C||A\cup\varepsilon|+|\beta||A\cup C|)s(\zeta,B))\\
&\qquad+|A||C|((|\alpha||A\cup C|+|\zeta||\alpha|+|\zeta||\delta\cup C|)s(B,\gamma)\\
&\qquad+|B||\delta\cup C|(s(B\setminus A,\gamma)+s(A\cap B,\gamma))+|B||\alpha|(s(\beta,\gamma)+s(\theta,\gamma))\\
&\qquad+(|A\cup\varepsilon||\alpha|+|A\cup\varepsilon||\delta\cup C|+|\beta||A\cup C|)s(B,\zeta))\\
&\qquad+|A||B|(|\alpha||A\cup\varepsilon|s(B\setminus C,C)+(|\alpha||\gamma|+|\zeta\cup B||A\setminus C|)s(\beta,C)\\
&\qquad+|\zeta\cup B||C|(s(\beta,A\cap C)+s(\beta,C\setminus A))+(|\beta||\alpha|+|A\cup B||\delta\cup C|)s(\delta,C))\\
&\qquad-|B||C|(|\zeta\cup B||\beta|s(A,C\setminus A)+(|\alpha||\beta|+|B\setminus A||\delta\cup C|)s(A,\gamma)\\
&\qquad+|A||\delta\cup C|(s(\alpha\cup\zeta,\gamma)+s(A\cap B,\gamma))+(|\beta||\gamma|+|A\cup\varepsilon||B\cup C|)s(A,\varepsilon))\\
&\qquad-|A||B|(|\beta||B\cup\zeta|s(A\setminus C,C)+(|\beta||\gamma|+|A\cup\varepsilon||B\setminus C|)s(\alpha,C)\\
&\qquad+|A\cup\varepsilon||C|(s(\alpha,\zeta\cup\gamma)+s(\alpha,B\cap C))+(|\alpha||\beta|+|A\cup B||\delta\cup C|)s(\delta,C))\\
&\quad=|B||C|(|\delta\cup C|t(A,B\setminus A,\gamma)+|\alpha|t(A,\beta,\gamma)+|B\cup\zeta|t(A,\beta,C\setminus A))\\
&\qquad+|A||B|(|A\cup\varepsilon|t(\alpha,B\setminus C,C)+|\gamma|t(\alpha,\beta,C)+|B\cup\zeta|t(A\setminus C,\beta,C))\\
&\qquad+|A||C|((|A\cup C|+|\zeta|)t(\alpha,B,\gamma)+|B|t(\alpha,\theta,\gamma))\\
&\qquad+|A||C|(|A\cup\varepsilon|t(\alpha,B,\zeta)+|C\cup\delta|t(\zeta,B,\gamma))\\
&\qquad+2|A||C|(|\delta\cup C||A\cup\varepsilon|+|\beta||A\cup C|)s(B,\zeta)\\
&\qquad+2|A||B||C||B\cup\zeta|s(\beta,A\cap C)\geq 0.
\end{aligned}
$$

where the equality holds if all terms of s and t are zero.

Appendix B: Triangle inequalities of Example 2

The triangle inequality for (17) can be proved as follows: Let $x=e^{p\lambda}$ and let

$$\tau(x)=\left(\frac{x|A\triangle B|+|A\cap B|}{|A\cup B|}\right)\left(\frac{x|B\triangle C|+|B\cap C|}{|B\cup C|}\right)-\left(\frac{x|A\triangle C|+|A\cap C|}{|A\cup C|}\right).$$

Then the triangle inequality for $p>0$ is equivalent to $\tau(x)\geq 0$ for $x>1$. The first derivative of $\tau(x)$ with respect to x is

$$\tau'(x)=2(x-1)j(A,B)j(B,C)+j(A,B)+j(B,C)-j(A,C),$$

where j is the Jaccard distance (4). Since $\tau(1) = 0$ and $\tau'(x) \geq 0$ for $x \geq 1$, we have $\tau(e^{p\lambda}) \geq 0$ for $p > 0$.

The triangle inequality for (18) can be proved as follows: let $y = 1 - e^{p\lambda}$ and let

$$\begin{aligned}\tau(y) &= \left(1 - \frac{|A \setminus C|}{|A \cup C|}y\right)\left(1 - \frac{|C \setminus A|}{|A \cup C|}y\right) \\ &\quad - \left(1 - \frac{|A \setminus B|}{|A \cup B|}y\right)\left(1 - \frac{|B \setminus A|}{|A \cup B|}y\right)\left(1 - \frac{|B \setminus C|}{|B \cup C|}y\right)\left(1 - \frac{|C \setminus B|}{|B \cup C|}y\right) \\ &= y\phi(y),\end{aligned}$$

where $\phi(y)$ is the cubic function of y with a negative leading coefficient. Then the triangle inequality for $p < 0$ is equivalent to $\tau(y) \geq 0$ for $y \in [0, 1)$. The function $\phi(y)$ satisfies the following inequalities: $\phi(0) \geq 0$, $\phi(1) = \tau(1) \geq 0$, $\phi'(1) \leq 0$, and $\phi''(1) \geq 0$. These inequalities can be proved by decomposition of A, B, and C into α, β, γ, δ, ε, ζ and η defined in Appendix A. Then, we have $\phi(y) \geq 0$ for $y \in [0, 1)$, therefore, $\tau(1 - e^{p\lambda}) \geq 0$ for $p < 0$.

References

1. Hart, K.P., Nagata, J., Vaughan, J.E. (eds.): Encyclopedia of General Topology. Elsevier, Amsterdam (2004)
2. Nagata, J.: Modern General Topology, 2nd rev. edn. North-Holland, Amsterdam (1985)
3. Rucklidge, W.: Efficient Visual Recognition Using the Hausdorff Distance. Lecture Notes in Computer Science, vol. 1173. Springer, Berlin (1996)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
5. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, London (2008)
6. Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer, Berlin (2009)
7. Bullen, P.S.: Handbook of Means and Their Inequalities. Mathematics and Its Applications, vol. 560. Kluwer, Dordrecht (2003)
8. Searcoid, M.O.: Metric Spaces. Springer, Berlin (2007)
9. Lowen, R.: Approach Spaces: The Missing Link in the Topology-Uniformity-Metric Triad. Oxford University Press, NY (1997)
10. Everitt, B.S.: Cluster Analysis. Heinemann, London (1980)
11. Bukatin, M., Kopperman, R., Matthews, S., Pajoohesh, H.: Partial Metric Spaces. Am. Math. Mon. **116**(8), 708–718 (2009)
12. Rubinstein, R.Y., Kroese, D.P.: Simulation and the Monte Carlo Method, 2nd edn. Wiley, New York (2008)
13. Zimmermann, H.J.: Fuzzy Set Theory and Its Applications, 4th edn. Kluwer, Dordrecht (2001)
14. IEEE Learning Technology Standard Committee: Learning object metadata. http://ltsc.ieee.org/wg12/
15. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
16. Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967)
17. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
18. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Haussler, D. (ed.) 5th Annual ACM Workshop on COLT, pp. 144–152. ACM Press, Pittsburgh (1992)